SYSTEMS DEVELOPMENT OFFICE
TECHNIQUES DEVELOPMENT LABORATORY

TDL OFFICE NOTE 78-15

AN OBJECTIVE METHOD FOR MAXIMIZING THREAT SCORE

Robert J. Bermowitz and Donald L. Best

December 1978

Robert J. Bermowitz and Donald L. Best

1. INTRODUCTION

Frequently, it is necessary to provide a categorical forecast of an event given only probabilities of several event categories. For example, at the Techniques Development Laboratory we make automated categorical forecasts of precipitation amount by comparing probability forecasts to a preselected threshold probability[1] that will maximize the threat score[2] for dichotomous forecasts of a category (Bermowitz and Zurndorfer, 1979). The purpose of this paper is to describe an objective method for determining threshold probabilities that is more efficient than the one currently used. In addition, we will describe the results of a verification in which forecasts made from threshold values obtained with this new method are compared to those obtained from the current technique.

The technique now used to compute threshold values which maximize the threat score is an empirical, iterative one. On successive passes through the dependent data sample, threat scores are computed for categorical forecasts made by comparing the actual probability forecasts against preselected, incremented threshold probabilities. The procedure is terminated when the threat score reaches its maximum value within the accuracy of the given increments. The threshold probability associated with that maximum threat score is then subjectively evaluated to see if it should be used operationally. Usually, this involves checking the bias[3] to make sure that it is not unacceptably high. If it is too high, a threshold value associated with a lower bias is chosen; unfortunately, this usually results in a lower threat score. Although the initial step to find the threshold probability which maximizes the threat score is certainly objective, the use of the accompanying bias information introduces subjectivity to the entire procedure.

---

1 The threshold probability for a category, say $\geq$ .25 inch of precipitation, is a value that if exceeded by a probability forecast for that category, would result in a categorical forecast of $\geq$ .25 inch. If the threshold value is not exceeded, the categorical forecast would be $<$ .25 inch.

2 Threat score = H/(F+O−H) where H is the number of correct forecasts of a category and F and O are, respectively, the number of forecasts and observations of that category.

3 Bias is the number of forecasts of a category divided by the number of observations of that category. A categorical bias equal to 1 means unbiased forecasts of that category.

at the top of the output is .21. The bias associated with this threshold probability, 1.81, is somewhat high. In this case, one may wish to choose a threshold probability of .23 which is associated with a lower bias (1.55) without much loss in threat score. Note, that there is a secondary maximum in the threat score at a threshold probability of .17; however, the bias is too high for this threshold to be used.

The important thing to note is that this iterative procedure, while adequate, is very time consuming from the standpoint of both the human and the computer. Consider that threshold probabilities must be evaluated for every category for all regions for all forecast periods. In a normal developmental effort for quantitative precipitation, there are 4 categories, 9 regions, and 12 projections, or 432 threshold probabilities to be obtained. Furthermore, prior to determining threshold probabilities, probability forecasts must be prepared by the forecast program, which consumes still more time. Therefore, it is obvious that it would be worthwhile to have a more efficient technique to determine threshold probabilities which could replace the current one without any loss in skill of the resulting forecasts.

2. DESCRIPTION OF METHOD

Generally, a high threshold probability is associated with an event which has a high relative frequency and vice versa. It is also true that the higher an event's relative frequency is, the more likely is an equation with a substantial reduction of variance to be found. This suggests a relationship between threshold probability and correlation coefficient, since higher forecast probabilities are obtained from regression equations with higher reductions of variance.

To better define this relationship, we plotted warm season, regionalized correlation coefficients for six different forecast projections obtained from our operational probability of precipitation amount (PoPA) equations against corresponding threshold probabilities. The latter were <u>not</u> determined through subjective evaluation; rather, they were chosen by the computer and were the ones that gave the highest threat score on the dependent data without regard to bias. The resulting plot, shown in Fig. 2, indicates that the relationship is quite linear for the range of threshold probabilities and correlation coefficients shown. The equation for the line of best fit, determined by least squares, is

$$T = -.028 + .597R,$$

where T is threshold probability and R is correlation coefficient. Here, we will refer to this equation as the R model. The reduction of variance in fitting T with the R value is a very good .912.

Recent work by Miller and Best (1978) to determine efficient methods of minimizing categorical bias (bias=1) has shown that the event climate (C) is an important parameter. Accordingly, we developed another regression equation in which R, C, and their product were introduced as predictors

The reduction of variance for this association is .947. The range of the predictand relative frequencies was from .0020 to .1612.

Finally, still another regression equation was developed in the form of the generalized threshold probability model discussed by Miller and Best (1978). This equation, which we will call the M&B model, is

$$T = .698R\,(.5-C) + C.$$

This equation produced a reduction of variance of .943.

To summarize, the new method we propose to maximize threat score computes a T value from either R alone or a combination of R and C by means of the R, RC, and M&B models (or equations). This method is far more efficient than the iterative technique since it requires no additional computer runs or subjective evaluations. In fact, the threshold probabilities can be computed at the completion of the regression program in which the forecast equations are developed.

3. VERIFICATION

To test the three models, which is tantamount to testing the new method, we performed a verification in which precipitation amount forecasts made from thresholds obtained from the three equations were compared to those made operationally from thresholds obtained from the iterative method. Four sets of independent data were available for the comparative verification: (1) 12-36 h forecasts from warm season (April-September), Primitive Equation (PE) (Shuman and Hovermale, 1968) model-based PoPA equations, (2) 12-36 h forecasts from cool season (October-March) PE-based PoPA equations, (3) 12-18 h cool season forecasts from Limited-area Fine Mesh (LFM) (Gerrity, 1977) model-based PoPA equations, and (4) 24-48 h cool season forecasts from PE-based PoPA equations for the Bonneville Power Administration (Bermowitz et al., 1977). The choice of these data sets represents, at least in part, an attempt to select independent data with as much difference from the dependent data as possible and with as much variety as possible. The 12-36 h cool season PE-based PoPA data consisted of two cool seasons; the other consisted of one season of data.

In all cases, threat score and biases were computed for forecasts of the precipitation amount categories $\geq$ .25, $\geq$ .50, $\geq$ 1.0, and $\geq$ 2.0 inches. An exception was the 12-18 h LFM-based PoPA data set for which forecasts of $\geq$ 2.0 inches were not available. Verification scores were computed at 233 cities over the conterminous U.S. for all data sets except Bonneville; these forecasts were made for and verified at 65 stations over the Columbia River Basin.

Comparative verifications on the four data sets are summarized in Tables 1-4. It is important to remember that we are not necessarily seeking a technique that improves upon the existing method in terms of verification. Since the new method is much more efficient than the existing one, we would be satisified if it gave results about as good as those obtained with the iterative technique.

Table 1 contains the results for the warm season, 12-36 h PE-based PoPA data. It can be seen that there is very little difference in threat scores among the three models. More importantly, there is very little difference between the models and the operational system. Overall, the R model has a somewhat lower bias than the others.

Results for the two cool seasons of 12-36 h PE-based PoPA data are given in Table 2. The threat score for the category $\geq$ 1.0 inch for the M&B model is lower than that of the R and RC models, while the latter are about as good as the operational system. Note, also, the excessively high bias for the M&B model for this category. All three models do not perform as well as the operational system for the category $\geq$ 2.0 inches (the only time this occurred). As shown by the bias, this may be due to the relative over-forecasting of this category by the operational system when compared to the three models. For the other two categories, the three models all perform at least as good as the operational system with the R and RC models having somewhat higher threat scores than the M&B model. Overall, the R and RC models have about the same threat scores, but the bias characteristics are better for the R model.

Table 3 contains the results for the cool season, 12-18 h LFM-based PoPA data. For the category $\geq$.25 inch, M&B has a somewhat lower threat score than the other models and the operational system. For all categories, the R and RC models have threat scores at least as good as the operational system; in fact, R has slightly better threat scores than either RC or the operational system. In addition, the bias for the R model is considerably better than those for the other models and the operational system.

Results for the cool season, 24-48 h Bonneville data are presented in Table 4. Threat scores for the M&B model are lowest of the group in all categories. Threat scores for the R and RC models are about the same and are at least as good as those of the operational system. Note that again R has a somewhat better bias than RC; overall, R's bias is about as good as that of the operational system.

Results for all four data sets were also broken down by region to determine if there were any poor regional threat scores masked by the overall results. With only one exception, there were none. In that case, (12-36 h cool season) the R model failed to produce any forecasts of the category $\geq$ 1.0 inch. The RC model, on the other hand, not only produced forecasts of that category but had 5 hits. Of particular interest is the fact that the threshold probability derived from the RC model was only .012 lower than that from the R model for the category $\geq$ 1.0 inch. The reason for this poor hold value can cause radical changes in the categorical forecast statistics. For example, the slightly lower threshold for RC with clustered forecasts allowed the category $\geq$ 1.0 inch to be forecast by RC but not by R. However, this problem is not serious because continuous predictors, which now are used with binaries, alleviate clustering.

We also examined how this new method will behave on data other than precipitation amount. In categorical thunderstorm forecasting, Foster and Reap (1978) have found through a procedure similar to the iterative method that a threshold probability of .350 maximizes the threat score of the category "occurrence of a thunderstorm." The threat score obtained with this threshold is .452. We used their correlation coefficient and climatology to compute threshold probabilities by means of the R, RC, and M&B models. These thresholds along with corresponding threat scores and biases are summarized in Table 5. It appears that the new method produces verification statistics about the same as those obtained by Foster and Reap; the M&B model is particularly close in this example.

Gofus (1978) has used the R model and the iterative technique (lack of a reliable climatology precluded use of the RC and M&B models) to compute threshold probabilities for categorical fog forecasting over the Great Lakes. Some preliminary results on independent data indicate that threat scores for the R model are nearly as good as those for the iterative technique.

5. SUMMARY AND CONCLUSIONS

A more efficient objective method for obtaining threshold probabilities that maximize the threat score has been presented. This method, when tested on independent precipitation amount data, gave results as good as those from thresholds obtained from the current iterative procedure and used operationally. (Strictly speaking, for the precipitation amount data, this was true for only the R and RC models.) In addition, there were indications that the new method can be used to maximize the threat score when forecasting events other than precipitation amount. For example, it appears to have held up when used in categorical thunderstorm and fog forecasting.

There is a question that remains concerning which model--R, RC, or M&B,--to use. For example, in forecasting precipitation amount, the R model performed the best. On the other hand, M&B appears to be the choice in forecasting thunderstorms. Perhaps the safest answer, therefore, is that potential users of this method do their own testing to determine which model to use. One thing that is certain, however, is that the R model is the only one that can be used if a reliable climatology is not available.

We strongly recommend that this new method be considered by those who require that their categorical forecasts produce a maximum threat score. The potential savings in both human and computer time over the current program has to be run to replace the ones which make the probability forecasts and perform the iterative procedure; threshold probabilities can be computed in the regression program used to develop the forecast equations.

6. REFERENCES

Bermowitz, R. J., and E. A. Zurndorfer, 1979: Automated guidance for predicting quantitative precipitation. Mon. Wea. Rev., (in press).

_____, _____, and J. P. Dallavalle, 1977: Development of updated cool season precipitation and temperature equations for the Columbia River Basin. Final Report, Phase IV. Prepared for the Department of Interior, Bonneville Power Administration, Portland, Oregon, 15 pp.

Foster, D. S., and R. M. Reap, 1978: Comparative verification of the operational 24-h convective outlooks with the objective severe local storm guidance based on model output statistics. TDL Office Note 78-7, National Weather Service, NOAA, U. S. Department of Commerce, 17 pp.

Gerrity, J. F., 1977: The LFM model - 1976: A documentation. NOAA Technical Memorandum NWS NMC-60, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 68 pp.

Gofus, J. G., 1978: Personal Communication.

Miller, R. G., and D. L. Best, 1978: A model for converting probability forecasts to categorical forecasts. TDL Office Note 78-14, National Weather Service, NOAA, U.S. Department of Commerce, 10 pp.

Shuman, F. G., and J. B. Hovermale, 1968: An operational six-layer primitive equation model. J. Appl. Meteor., 7, 525-547.

THREAT SCORE HAS BEEN MAXIMIZED—THRESHOLD PROBABILITY FOR THE CATEGORY $\geq$ 0.25 = 0.210

ON	THREAT SCORE	PERCENT CORRECT	SKILL SCORE	POST AGREEMENT	PREFIGURANCE	BIAS	THD.
	0.245	79.8	0.308	0.26	0.80	3.05	0.
	0.253	81.1	0.321	0.27	0.78	2.86	0.
	0.261	82.2	0.334	0.28	0.77	2.70	0.
	0.271	83.4	0.350	0.30	0.75	2.53	0.
	0.273	84.1	0.354	0.33	0.73	2.39	0.
	0.269	84.7	0.351	0.31	0.69	2.24	0.
	0.267	85.4	0.349	0.31	0.65	2.07	0.
	0.272	86.3	0.358	0.33	0.62	1.92	0.
	0.274	86.9	0.363	0.33	0.60	1.81	0.
	0.268	87.3	0.357	0.34	0.57	1.68	0.
	0.267	87.9	0.357	0.35	0.54	1.55	0.
	0.262	88.4	0.352	0.35	0.50	1.42	0.
	0.263	89.0	0.357	0.37	0.48	1.29	0.
	0.260	89.4	0.355	0.38	0.45	1.20	0.
	0.259	89.8	0.356	0.39	0.43	1.10	0.
	0.250	90.1	0.346	0.40	0.40	1.01	0.
	0.242	90.4	0.337	0.41	0.37	0.92	0.
	0.236	90.8	0.333	0.42	0.35	0.82	0.
	0.232	91.1	0.330	0.44	0.33	0.74	0.
	0.220	91.2	0.316	0.45	0.30	0.67	0.

1. An example of computer output from the iterative procedure which selects a threshold probability
mize the threat score.

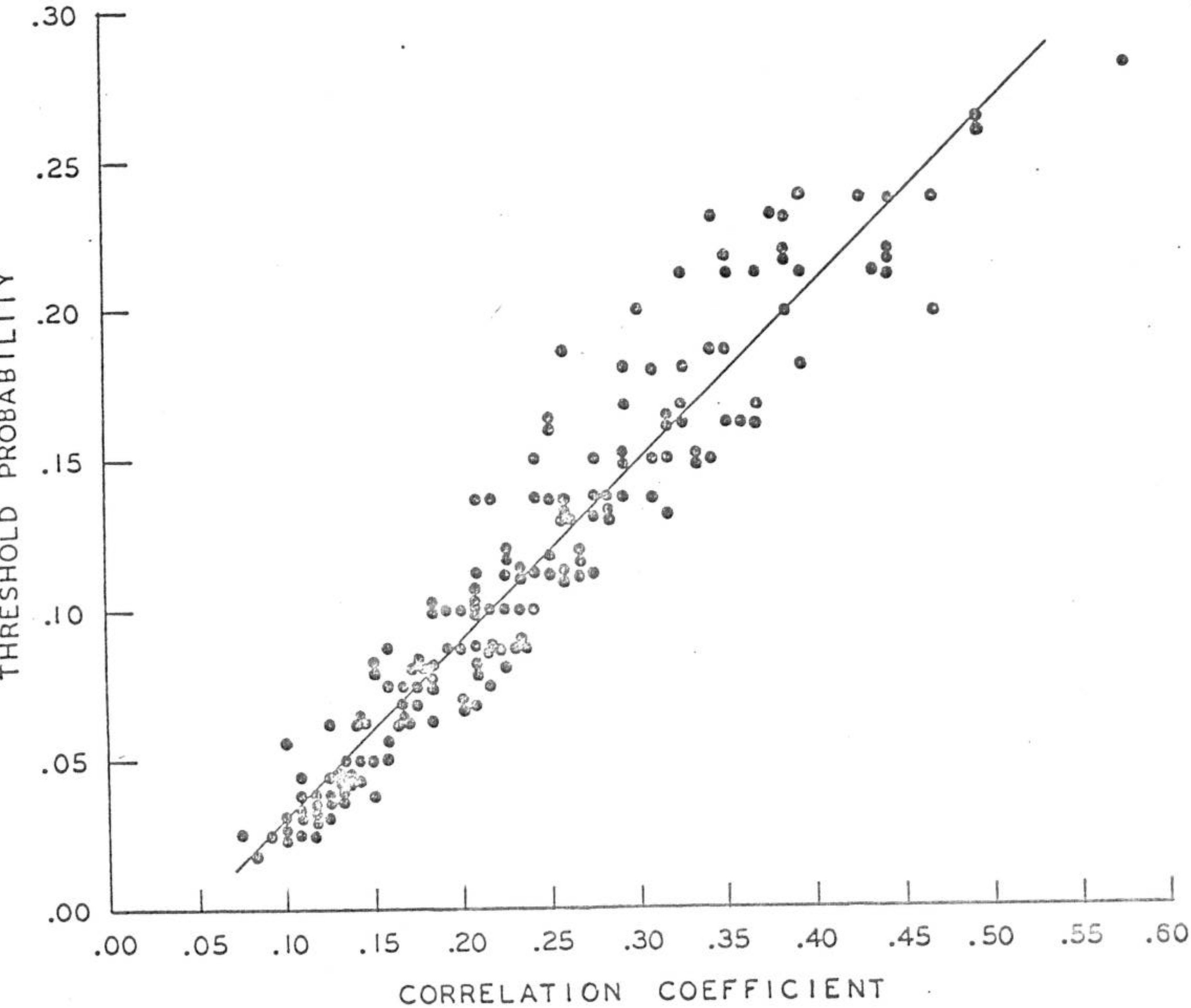

Figure 2. A plot of threshold probabilities against correlation coefficients for warm season (April-September) probability of precipitation amount data for six different forecast projections. Sample size equals 166.

September) 12-based PoPf categorical forecasts made from threshold probabilities from the (1) iterative method (OPER), (2) R, (3) RC, and (4) M&B models. Sample consists of one season of forecasts at 233 cities.

CATEGORY (INCH)	THREAT SCORE				BIAS			
	OPER	R	RC	M&B	OPER	R	RC	M&B
$\geq$.25	.250	.245	.248	.249	1.88	2.03	1.94	1.76
$\geq$.50	.175	.179	.178	.180	1.91	1.91	1.84	1.86
$\geq$1.0	.093	.094	.095	.096	1.78	1.63	1.66	1.87
$\geq$2.0	.024	.029	.029	.033	1.60	1.21	1.98	1.92

Table 2. Same as Table 1 except for two cool seasons (October-March of data.

CATEGORY (INCH)	THREAT SCORE				BIAS			
	OPER	R	RC	M&B	OPER	R	RC	M&B
$\geq$.25	.313	.338	.325	.316	1.88	1.65	1.93	1.94
$\geq$.50	.229	.246	.233	.231	2.41	2.17	2.47	2.57
$\geq$1.0	.161	.152	.164	.124	2.07	1.79	2.29	3.99
$\geq$2.0	.054	.013	.015	.021	2.20	0.76	0.94	1.16

Table 3. Same as Table 1 except for one cool season of 12-18 h LFM-based PoPA categorical forecasts.

CATEGORY	THREAT SCORE				BIAS			
(INCH)	OPER	R	RC	M&B	OPER	R	RC	M&B
$\geq$.25	.265	.274	.272	.257	1.68	1.30	1.49	1.86
$\geq$.50	.175	.180	.176	.178	1.95	1.41	1.72	2.33
$\geq$1.0	.080	.088	.081	.080	2.65	1.32	2.14	2.64
$\geq$2.0	--	--	--	--	--	--	--	--

Table 4. Same as Table 1 except for one cool season of 24-48 h forecasts for stations over the Columbia River Basin.

CATEGORY	THREAT SCORE				BIAS			
(INCH)	OPER	R	RC	M&B	OPER	R	RC	M&B
$\geq$.25	.411	.406	.410	.401	1.42	1.59	1.63	1.34
$\geq$.50	.364	.371	.366	.358	1.50	1.63	1.73	1.51
$\geq$1.0	.255	.259	.251	.241	1.81	1.77	1.88	1.90
$\geq$2.0	.162	.171	.173	.147	1.46	1.07	1.26	1.68

Table 5. Comparison of threshold probabilities that maximize the threat score for categorical thunderstorm forecasting, corresponding threat scores, and categorical biases for the (1) iterative method (OPER), (2) R, (3) RC, and (4) M&B models.

MODEL	THRESH. PROB.	THREAT SCORE	BIAS
OPER	.350	.452	1.45
R	.312	.450	1.63
RC	.284	.444	1.76
M&B	.359	.451	1.41